Εργασία Α8.

Σκοπός της Εργασίας

Η παρούσα εργασία είναι μέρος του Παραδοτέου Π1. της Ροής Α και περιγράφει αναλυτικά το τρόπο δημιουργίας, επικαιροποίησης και συντήρησης των μητρώων απο το σύστημα και τις εξωτερικές πηγές εμπλουτισμού/καθαρισμού.

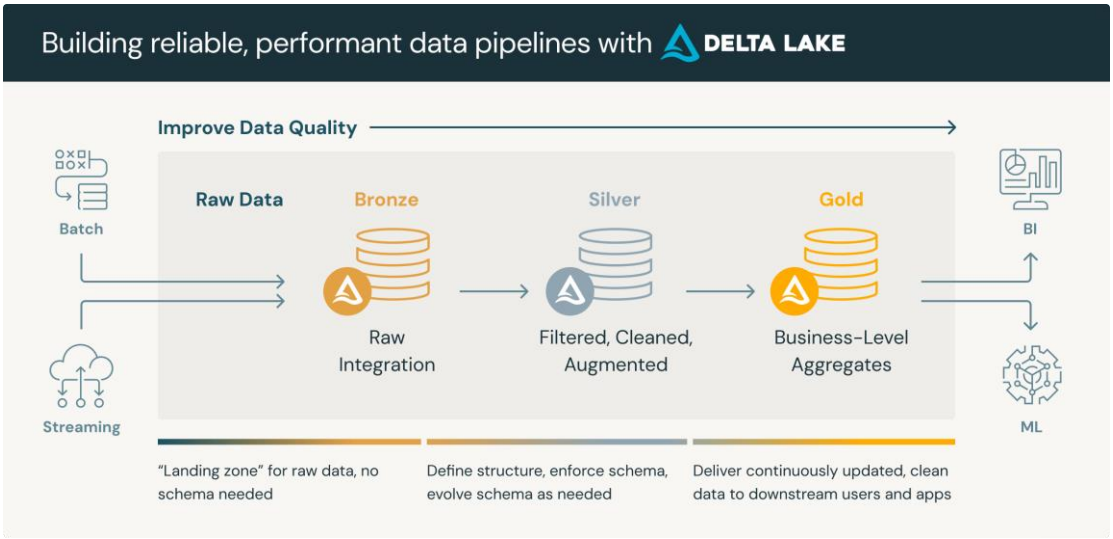
Τεχνικό Σχέδιο για τη Δημιουργία, Εισαγωγή και Συντήρηση Πινάκων Delta στο Databricks

1. Εισαγωγή

Ο στόχος αυτού του εγγράφου είναι να περιγράψει τον τρόπο δημιουργίας, εισαγωγής και συντήρησης πινάκων Delta στο Databricks για τη διαχείριση δεδομένων που προέρχονται από τις οντότητες των μητρώων που περιλαμβάνονται στις απαιτήσεις του συστήματος.

2. Αρχιτεκτονική Συστήματος

Η αρχιτεκτονική medallion είναι ένα μοτίβο σχεδιασμού δεδομένων που χρησιμοποιείται για τη λογική οργάνωση των δεδομένων σε ένα lakehouse , με στόχο τη σταδιακή και προοδευτική βελτίωση της δομής και της ποιότητας των δεδομένων καθώς ρέουν μέσα από κάθε στρώμα της αρχιτεκτονικής (από πίνακες επιπέδων Bronze ⇒ Silver ⇒ Gold) .



Το Databricks παρέχει εργαλεία όπως το Delta Live Tables (DLT) που επιτρέπουν στους χρήστες να δημιουργούν άμεσα αγωγούς δεδομένων με πίνακες Bronze, Silver

και Gold από λίγες μόνο γραμμές κώδικα. Και, με πίνακες ροής και υλοποιημένες προβολές, οι χρήστες μπορούν να δημιουργήσουν αγωγούς ροής DLT που βασίζονται στο Apache Spark™ Structured Streaming που ανανεώνονται και ενημερώνονται σταδιακά.

Bronze Layer (ακατέργαστα δεδομένα)

Το στρώμα Bronze είναι όπου προσγειώνουμε όλα τα δεδομένα από εξωτερικά συστήματα πηγών. Οι δομές του πίνακα σε αυτό το επίπεδο αντιστοιχούν στις δομές πίνακα του συστήματος πηγής "ως έχουν", μαζί με τυχόν πρόσθετες στήλες μεταδεδομένων που καταγράφουν την ημερομηνία/ώρα φόρτωσης, το αναγνωριστικό διεργασίας κ.λπ. Η εστίαση σε αυτό το επίπεδο είναι η γρήγορη αλλαγή της σύλληψης δεδομένων και η δυνατότητα παροχής ιστορικού αρχείου πηγής (ψυχρή αποθήκευση), γενεαλογία δεδομένων, δυνατότητα ελέγχου, επανεπεξεργασία εάν χρειάζεται χωρίς να ξαναδιαβαστούν τα δεδομένα από το σύστημα προέλευσης.

Silver Layer (καθαρισμένα και συμμορφωμένα δεδομένα)

Στο στρώμα Silver του lakehouse, τα δεδομένα από το στρώμα Bronze αντιστοιχίζονται, συγχωνεύονται, συμμορφώνονται και καθαρίζονται ("ακριβώς-αρκετά") έτσι ώστε το στρώμα Silver να παρέχει μια "Επιχειρηματική προβολή" όλων των βασικών επιχειρηματικών οντοτήτων, εννοιών και συναλλαγές. (π.χ. κύριοι πελάτες, καταστήματα, μη διπλότυπες συναλλαγές και πίνακες παραπομπών).
Το επίπεδο Silver φέρνει τα δεδομένα από διαφορετικές πηγές σε μια προβολή Enterprise και επιτρέπει την ανάλυση αυτοεξυπηρέτησης για ad-hoc αναφορές, προηγμένα αναλυτικά στοιχεία και ML. Χρησιμεύει ως πηγή για τους Αναλυτές Τμήματος, τους Μηχανικούς Δεδομένων και τους Επιστήμονες Δεδομένων για περαιτέρω δημιουργία έργων και αναλύσεων για την απάντηση επιχειρηματικών προβλημάτων μέσω έργων δεδομένων επιχειρήσεων και τμημάτων στο Gold Layer.

Στο παράδειγμα μηχανικής δεδομένων lakehouse, συνήθως ακολουθείται η μεθοδολογία ELT έναντι ETL - που σημαίνει ότι εφαρμόζονται μόνο ελάχιστοι ή «αρκετοί» μετασχηματισμοί και κανόνες καθαρισμού δεδομένων κατά τη φόρτωση του επιπέδου Silver. Η ταχύτητα και η ευελιξία για την απορρόφηση και παράδοση των δεδομένων στη λίμνη δεδομένων έχουν προτεραιότητα και εφαρμόζονται πολλοί σύνθετοι μετασχηματισμοί και επιχειρηματικοί κανόνες για συγκεκριμένο έργο κατά τη φόρτωση των δεδομένων από το επίπεδο Silver σε Gold. Από την άποψη της μοντελοποίησης δεδομένων, το Silver Layer έχει περισσότερα μοντέλα δεδομένων 3ης-κανονικής μορφής.

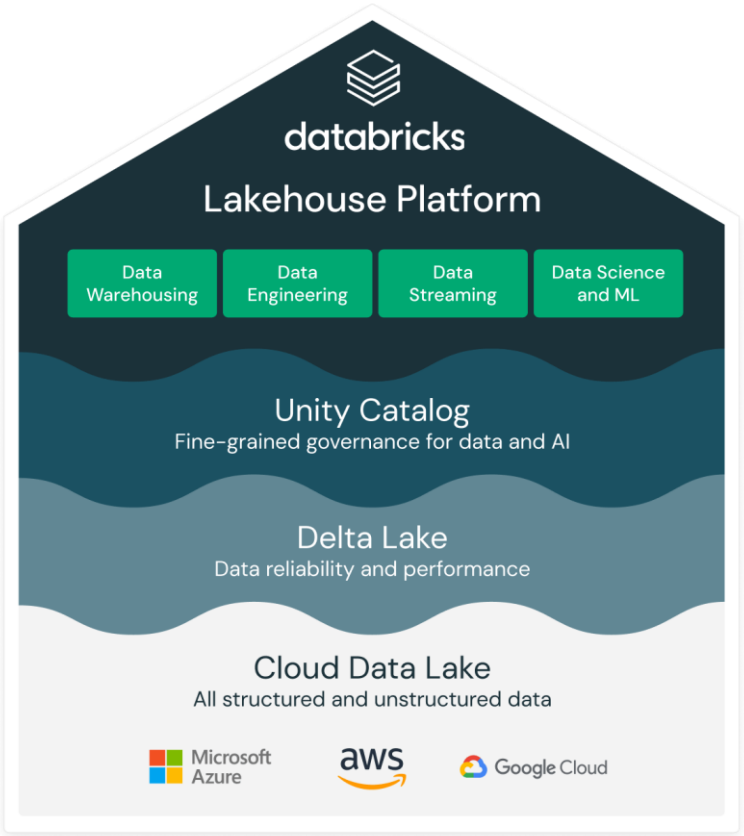
Gold Layer (Curated business data)

Τα δεδομένα στο στρώμα Gold του lakehouse είναι συνήθως οργανωμένα σε βάσεις δεδομένων «ειδικών για το έργο» έτοιμες για κατανάλωση. Το επίπεδο Gold προορίζεται για αναφορές και χρησιμοποιεί περισσότερα αποκανονικοποιημένα και βελτιστοποιημένα για ανάγνωση μοντέλα δεδομένων με λιγότερες συνδέσεις. Το τελικό επίπεδο των μετασχηματισμών δεδομένων και οι κανόνες ποιότητας δεδομένων εφαρμόζονται εδώ. Το τελικό επίπεδο παρουσίασης των μητρώων που περιλαμβάνονται στο σύστημα θα δημιουργείτε σε αυτό το επίπεδο.

Εδώ, μπορεί να φανεί πως τα δεδομένα επιμελούνται καθώς κινούνται μέσα από τα διαφορετικά στρώματα ενός Lakehouse. Σε ορισμένες περιπτώσεις, είναι δυνατή επίσης η εισαγωγή πολλών Data Marts και EDW από την παραδοσιακές RDBMS στο lakehouse.

Πλεονεκτήματα του lakehouse architecture

- Simple data model
- Easy to understand and implement
- Enables incremental ETL
- Can recreate your tables from raw data at any time
- ACID transactions, time travel



3. Στοιχεία Συστήματος

- **Azure Synapse**: Περιβάλλον για το κεντρικό έλεγχο των ροών δεδομένων.
- **Databricks Workspace**: Περιβάλλον για την ανάπτυξη και διαχείριση big data εφαρμογών.
- **Spark Engine**: Χρησιμοποιείται για την επεξεργασία δεδομένων και τη δημιουργία πινάκων Delta.
- **Azure Data Lake Storage (ADLS)**: Αποθήκευση αρχείων δεδομένων.
- **Delta Lake**: Αρχιτεκτονική αποθήκευσης δεδομένων που παρέχει αξιόπιστη αποθήκευση και αποκατάσταση.

Η εισαγωγή δεδομένων περιλαμβάνει τη διαδικασία εισαγωγής δεδομένων για άμεση χρήση ή αποθήκευση σε μια βάση δεδομένων. Τα ακόλουθα βήματα περιγράφουν τη διαδικασία:

Πηγές Δεδομένων

- **Συστήματα Πηγής**: Τα δεδομένα μπορεί να προέρχονται από διάφορα συστήματα πηγής όπως συστήματα διαχείρισης πελατών, ERP, CRM, κ.λπ.
- **Μορφές Αρχείων**: Τα δεδομένα μπορεί να είναι σε διάφορες μορφές όπως CSV, JSON, XML κ.λπ.

Μέθοδοι Εισαγωγής

- **Batch Processing**: Εισαγωγή μεγάλων παρτίδων δεδομένων σε συγκεκριμένα χρονικά διαστήματα.
- **Streaming**: Συνεχής εισαγωγή δεδομένων σε πραγματικό χρόνο.
- **API Integration**: Χρήση API για την εισαγωγή δεδομένων από εξωτερικά συστήματα σε πραγματικό χρόνο ή ρονοπρογραμματισμένα.

Ροή Δεδομένων

- **Εισαγωγή Δεδομένων**: Τα δεδομένα φορτώνονται μέσω του Synapse Integration runtime σε κατάλληλες δομές δεδομένων σε storage account.
- **Μετασχηματισμός Δεδομένων**: Χρησιμοποιώντας Spark, τα δεδομένα μετασχηματίζονται σε κατάλληλη μορφή για να αποθηκευτούν ως Delta table.
- **Δημιουργία Πίνακα Delta**: Τα μετασχηματισμένα δεδομένα αποθηκεύονται σε έναν πίνακα Delta.
- **Συντήρηση Πίνακα**: Περιλαμβάνει διαδικασίες όπως ανανέωση δεδομένων, διαγραφή παλιών δεδομένων, και βελτιστοποίηση.

4. Σχεδιασμός Λύσης

a. Δημιουργία Πίνακα Delta

i. **Φόρτωση Δεδομένων από ADLS**:
   - Χρησιμοποιώντας τη βιβλιοθήκη Spark, τα δεδομένα φορτώνονται απο το Bronze Layer

ii. **Μετασχηματισμός Δεδομένων**:
   - Καθαρισμός και μορφοποίηση δεδομένων για τη δημιουργία ενός δομημένου πίνακα στο Silver Layer

iii. **Αποθήκευση Δεδομένων ως Delta Table**:
   - Χρησιμοποιώντας τις δυνατότητες της Delta Lake, τα δεδομένα αποθηκεύονται σε έναν πίνακα Delta στο Gold Layer.

b. Εισαγωγή Δεδομένων

- **Batch Ingestion**: Τα δεδομένα εισάγονται σε μεγάλες όγκους καθημερινά ή και ωριαίαμέσω του self-hosted integration runtime
- **Streaming Ingestion**: Σε περίπτωση συνεχούς ροής δεδομένων, χρησιμοποιούνται Spark Structured Streaming για την αποθήκευση των δεδομένων σε πραγματικό χρόνο.

c. Συντήρηση Πίνακα Delta

- **Ανανέωση Δεδομένων**: Χρησιμοποιώντας Spark Structured Streaming ή batch jobs για την ανανέωση των δεδομένων.
- **Διαγραφή Παλιών Δεδομένων**: Χρησιμοποιώντας τις δυνατότητες της Delta Lake για την αφαίρεση παλιών ή μη χρησιμοποιούμενων δεδομένων.
- **Βελτιστοποίηση**: Εκτέλεση των εντολών `OPTIMIZE` και `VACUUM` για τη βελτίωση της απόδοσης.

5. Συμπεράσματα

Αυτό το τεχνικό σχέδιο παρέχει τις οδηγίες για τη δημιουργία, εισαγωγή και συντήρηση πινάκων Delta στο Databricks. Η χρήση της Delta Lake εξασφαλίζει τη σταθερότητα και την αποτελεσματική διαχείριση των δεδομένων, βελτιώνοντας την απόδοση και την ανθεκτικότητα της υποδομής δεδομένων.